Capturing Complex Spatio-Temporal Relations among Facial Muscles for Facial Expression Recognition

Ziheng Wang[1] Shangfei Wang[2] Qiang Ji[1]
[1]ECSE Department, Rensselaer Polytechnic Institute
[2]School of Computer Science and Technology, University of Science and Technology of China
{wangz10,jiq}@rpi.edu sfwang@ustc.edu.cn

Abstract

Spatial-temporal relations among facial muscles carry crucial information about facial expressions yet have not been thoroughly exploited. One contributing factor for this is the limited ability of the current dynamic models in capturing complex spatial and temporal relations. Existing dynamic models can only capture simple local temporal relations among sequential events, or lack the ability for incorporating uncertainties. To overcome these limitations and take full advantage of the spatio-temporal information, we propose to model the facial expression as a complex activity that consists of temporally overlapping or sequential primitive facial events. We further propose the Interval Temporal Bayesian Network to capture these complex temporal relations among primitive facial events for facial expression modeling and recognition. Experimental results on benchmark databases demonstrate the feasibility of the proposed approach in recognizing facial expressions based purely on spatio-temporal relations among facial muscles, as well as its advantage over the existing methods.

1. Introduction

Facial expressions are the outcome of a set of muscle motions over a time interval. These movements interact in different patterns and convey different expressions. Understanding such complex facial activity not only requires us to study each individual facial muscle motion, but also how they interact with each other in both the space and time domain. Spatially, facial muscle motions can co-occur or can be mutually exclusive at each time slice. Temporally, the movement of one facial muscle can activate, overlap or follow another muscle. These spatio-temporal relations capture significant information about facial expressions yet have not been thoroughly studied, partially due to the limitations of the current models. Unlike most of the existing works that perform facial expression recognition on the manually labeled peak frame, we model a facial expression as a complex activity that spans over a time interval and consists of a group of primitive facial events happening sequentially or in parallel. More importantly, modeling facial expression as such a complex activity allows us to further study and capture a larger variety of complex spatial and temporal interactions among the primitive events. In this work, we aim to overcome the limitations of current models and thoroughly explore and exploit more complex spatio-temporal relations in the facial activities for expression recognition.

Understanding a complex activity and capturing the underlying temporal relations is challenging and most of the existing methods do not handle this adroitly. Modeling and recognizing a complex activity is naturally solved by building a structure that is able to semantically capture the spatio-temporal relationships among primitive events. Among various visual recognition methodologies, such as graphical, syntactic and description-based approaches, time-sliced graphical models, i.e. hidden Markov models (HMMs) and dynamic Bayesian networks (DBNs), have become the most popular tool for modeling and understanding complex activities [2, 11, 13, 5]. Syntactic and description-based approaches have also gained attention and used mainly for action units detection in recent years [10]. While these approaches have been applied to capture the dynamics of facial expressions, they face one or more of the following issues when modeling and understanding complex visual activities that involve interactions between different entities over durations of time.

First, time-sliced (based on time points) graphical models (e.g. HMM, DBN, or their variants) typically represent an activity as a sequence of instantaneously occurring events, which is generally unrealistic for facial expression. For example, the eye movement and nose movement may both last for a period of time and they may overlap. More-

over, time-sliced dynamic models can only offer three time-point relations (precedes, follows, equals), and so they are not expressive enough to capture many of the temporal relations between events that happen over the duration of an activity. Secondly, time-sliced graphical models typically assume first order Markov property and stationary transition. Hence, they can only capture local stationary dynamics and cannot represent global temporal relations. Finally, syntactic and description-based models lack the expressive power to capture and propagate the uncertainties associated with event detection and with their temporal dependencies in a principled manner.

To address these issues and comprehensively model facial expression, we propose a unified probabilistic framework that combines the probabilistic semantics of Bayesian networks (BNs) with the temporal semantics of interval algebra (IA). Termed an interval temporal Bayesian network (ITBN), this framework employs the BN's probabilistic basis and the IA's temporal relational basis in a principled manner that allows us to represent not only the spatial dependences among the primitive facial events but also a larger variety of time-constrained relations, while remaining fully probabilistic and expressive of uncertainty. In particular, ITBN is time-interval based in contrast to time-sliced models, which allows us to model the relations among both sequential and overlapping temporal events. In this paper we take a holistic approach to modeling the facial activities. We will first identify all of the related primitive facial events, which provide us the basis to define a larger variety of temporal relations. We then apply ITBN to capture their spatio-temporal interactions for expression recognition.

The remainder of this paper is organized as follows. Section 2 presents an overview of the related works. Section 3 introduces the definition and implementation of ITBN. We discuss how we identify the primitive facial events and how we model the facial expressions with ITBN in Section 4. Experiments and the discussions will be illustrated in Section 5. The paper is concluded in Section 6.

2. Related Works

Recognizing facial expressions generally involves bottom-level feature extraction and top-level classifier design. Features for expression classification can be grouped into appearance features such as Gabor [8] and LBP [9], and geometric features that are extracted from the location of the salient facial points. While appearance features capture the local or global appearance information of the facial components, studying the movement of the facial feature points provides us a more explicit manner to analyze the dynamics. Classifiers for facial expression recognition include static models and dynamic models. Static models recognize facial expressions based on the apex frame of an image sequence and have achieved successful performance. However, peak frames usually require manual labeling and the static approach completely disregards the dynamic interactions among the facial muscles that are very important for discriminating facial activates. In contrast dynamic models rely on the whole image sequence and study their temporal dynamics for facial expression recognition. In this paper we focus on expression recognition works that are based on the facial feature points and image sequences. A more comprehensive literature review of facial expression recognition ban be found in [14].

Dynamic models that have been widely applied for facial expression recognition include the hidden Markov model (HMM) and its variants [2, 11, 13], the dynamic Bayesian network (DBN) [5], and latent conditional random fields (LCRF) [4]. HMM captures local state transitions that are assumed to be stationary. In [2] a multilevel HMM is introduced to automatically segment and recognize human facial expressions from image sequences based on the local deformations of the facial feature points tracked with a piecewise Bezier volume deformation tracker. In [11], a non-parametric discriminant HMM is applied to recognize the expression and the facial features are tracked with Active Shape Model. A different approach is used in [13], where an HMM was used together with support vector machines and AdaBoost to simultaneously recognize action units and facial expressions by modeling the dynamics among the action units. Similarly, DBN also captures local temporal interactions and an example can be found in [5]. Besides these generative models, discriminative approaches such as LCRF have also been applied for expression analysis. For instance, in [4], features from 68 landmark points of video sequences were fed into an LCRF to perform expression recognition. However, all of these models are time-slice based and as a result can only capture a small portion of the temporal relations. Moreover, these relations are assumed to be stationary and time-independent. Therefore the captured dynamics remain local. To overcome these restrictions our proposed method models a complex activity as sequential or overlapping primitive events, and each event spans over a time interval. This allows us to capture a wider variety of complex temporal relations which can further enhance the performance of facial activity recognition.

3. Interval Algebra Bayesian Network

Different spatial and temporal configurations of primitive facial events lead to different expressions. Unlike the related works, ITBN looks at facial activity from a global view and is able to model a larger variety of spatio-temporal relations. To formally introduce the definition of ITBN, we will first define the primitive events that constitute a complex activity and several related concepts. A primitive event is also called a temporal entity and we do not differentiate

between these two terms in the remainder of this paper. We then introduce how we model the temporal relations among primitive events. Finally, we will formally introduce ITBN and its implementation.

Definition 1 (Temporal Entity) *A temporal entity is characterized by a pair* $\langle \Sigma, \Omega \rangle$ *in which* Σ *is a set of all possible states for the temporal entity, and* $\Omega = \{[a, b] \subset \mathbb{R} : a < b\}$ *is a period of time spanned by the temporal entity, where* a *and* b *denote the start time and the end time, respectively.*

Temporal entities form the primitive events of a complex activity. Spatio-temporal relations act as the joints connecting the temporal entities to form different patterns.

Definition 2 (Temporal Reference) *If a temporal entity* X *is used as a time reference for specifying temporal relations to another temporal entity* Y*, then* X *is the temporal reference of* Y*.*

Definition 3 (Temporal Dependency) *A temporal dependency (TD) denoted as* $I_{X,Y}$ *describes a temporal relation between two temporal entities* $X = \langle \Sigma_X, \Omega_X \rangle$ *and* $Y = \langle \Sigma_Y, \Omega_Y \rangle$*, where* X *is the temporal reference of* Y*.*

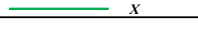
Relation	Symbol	Inverse	Illustration
Y before X	b	bi	
Y meets X	m	mi	
Y overlaps X	o	oi	
Y starts X	d	si	
Y during X	f	di	
Y finishes X	g	fi	
Y equals X	eq	eq	

Figure 1: Temporal Relations

Following Allen's Interval Algebra [1], there are a total of 13 temporal relationships between two temporal entities as illustrated in Figure 1. The thirteen possible relations $\mathbb{I} = \{\text{b, bi, m, mi, o, oi, s, si, d, di, f, fi, eq}\}$ respectively represent before, meets, overlaps, starts, during, finishes, equals and their inverses. The horizontal bars represent the time interval of the corresponding temporal entity. Given X and Y with X serving as the temporal reference, the dependency or temporal relation can be uniquely ascertained by the interval distance between two temporal entities as defined in Equation 1, where t_{xs} and t_{xe} (t_{ys} and t_{ye}) represent the start and end time of X (Y). Table 1 shows how we map the temporal distance to the temporal relationship.

$$d(X, Y) = [t_{xs} - t_{ys}, t_{xs} - t_{ye}, t_{xe} - t_{ys}, t_{xe} - t_{ye}] \quad (1)$$

The temporal dependency I_{XY} is graphically represented as a directed link leading from the node X to the node Y labeled with $I_{XY} \in \mathbb{I}$, as shown in Figure 2a. The

Table 1: Interval relation determined by interval distance

No.	r	$t_{xs} - t_{ys}$	$t_{xe} - t_{ye}$	$t_{xs} - t_{ye}$	$t_{xe} - t_{ys}$
1	b	< 0	< 0	< 0	< 0
2	bi	> 0	> 0	> 0	> 0
3	d	> 0	< 0	< 0	> 0
4	di	< 0	> 0	< 0	> 0
5	o	< 0	< 0	< 0	> 0
6	oi	> 0	> 0	< 0	> 0
7	m	< 0	< 0	< 0	$= 0$
8	mi	> 0	> 0	$= 0$	> 0
9	s	$= 0$	< 0	< 0	> 0
10	si	$= 0$	> 0	< 0	> 0
11	f	> 0	$= 0$	< 0	> 0
12	fi	< 0	$= 0$	< 0	> 0
13	eq	$= 0$	$= 0$	-	-

strength of the temporal dependency can be quantified by a conditional probability as follows:

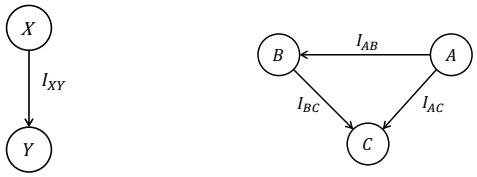
$$P(I_{XY} = i | X = x, Y = y), \quad (2)$$

where $x \in \Sigma_X$ and $y \in \Sigma_Y$ are the states of the temporal entities and $i \in \mathbb{I}$ denotes an interval temporal relation. Here, we only consider pairwise temporal dependencies.

Given the above concepts, we can formally introduce the ITBN as follows.

Definition 4 (Interval Temporal Bayesian Network) *An interval temporal Bayesian network (ITBN) is a directed graph (*DAG*)* $G(V, E)$*, where* V *is a set of nodes representing temporal entities and* E *is a set of links representing both the spatial and temporal dependencies among the temporal entities in* V*.*

A link in an ITBN is a carrier of the interval temporal relationship, and the link direction leading from X to Y indicates Y is temporally dependent on X and X is the temporal reference of Y. Once the temporal reference is established, the direction of the arc cannot be changed. It can only point from the temporal reference to the other temporal entity, thereby avoiding temporal relationship ambiguity. An example of ITBN can be seen in Figure 2b, which contains three temporal entities: A, B and C.



(a) **(b)**

Figure 2: (a) Graphical notation of the temporal dependency between primitive events X and Y. (b) An example of ITBN.

We propose to implement ITBNs with a corresponding Bayesian network (BN) to exploit the well-developed BN mathematical machinery. Figure 3 is the corresponding BN graphical representation for the ITBN shown in Figure 2b, where another set of nodes (the square nodes) are introduced to represent the temporal relations. Specifically, an ITBN implemented as a BN, includes two types of nodes:

temporal entity nodes (circular) and temporal relation nodes (square). There are also two types of links, spatial links (solid lines) and temporal links (dotted lines). The spatial links connect temporal entity nodes and capture the spatial dependencies among the temporal entities. The temporal links connect the temporal relation nodes with the corresponding temporal entities and characterize the temporal relationships between the two connected temporal entities. Given this representation, the joint probability of the temporal entities as well as their spatial and temporal information can be calculated with Equation 3:

$$P(\mathcal{Y}, \mathcal{I}) = \prod_{j}^{n} P(Y_j|\pi(Y_j)) \prod_{k}^{K} P(I_k|\pi(I_k)) \quad (3)$$

where $\mathcal{Y} = \{Y_j\}_{j=1}^{n}$ and $\mathcal{I} = \{I_k\}_{k=1}^{K}$ represent all temporal entity nodes and all temporal relation nodes respectively in an ITBN. $\pi(Y_j)$ is the set of parental nodes of Y_j; I_k represents the k^{th} interval temporal relation node and $\pi(I_k)$ are the two temporal entity nodes that produce I_k.

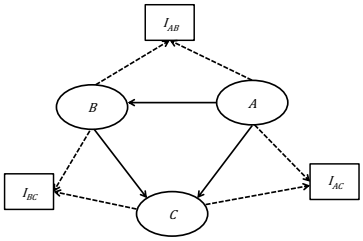


Figure 3: BN Implementation of ITBN

4. Facial Expression Recognition with ITBN

ITBN provides a powerful tool to model complex activities such as facial expression that consists of interval-based primitive temporal entities or events and captures a wider variety of spatial and temporal relations among them. In this section we will introduce the definitions of the primitive events that constitute a facial activity and how we capture their spatial and temporal relations with ITBN.

4.1. Facial Expression Modeling

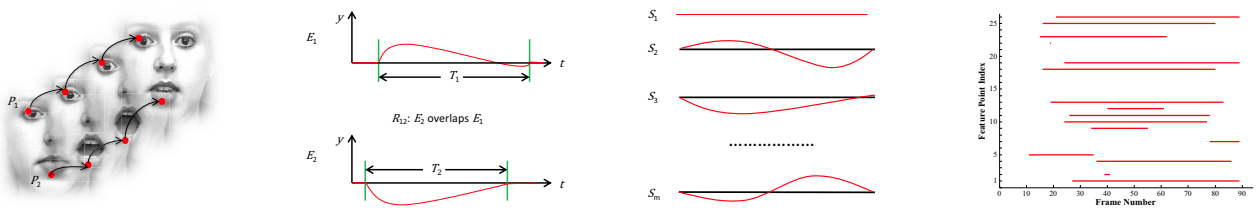
To comprehensively incorporate different levels of information from facial expression, the first step is to identify all of the related primitive facial events that constitute a facial activity. Primitive events for facial expressions are defined as the local facial muscle movements. Due to the difficulty of measuring facial muscle motions, we propose to approximate them using the movements of facial feature points. Facial feature points near different facial components are tracked and their movements are the result of different facial muscles (see Figure 6 for the facial feature points in our experiments). Therefore, each primitive facial event is defined as the movement of one facial feature point. Figure 4a shows two primitive facial events. Facial feature point P_1 corresponds to event E_1 and represents the movement of one of the left eye muscles. Point P_2 corresponds to event E_2 and captures the movement of one of the mouth muscles. A primitive facial event includes its temporal duration and state. In our case, the duration of the primitive event E_i starts when point P_i leaves its neutral position and ends at the time when it finally comes back. Basically, it is the time interval that point i stays away from its neutral position as shown in Figure 4b in which T_1 and T_2 are the corresponding duration for E_1 and E_2. For simplicity, only the trace along the vertical direction is shown in Figure 4. Each event has m possible states, which represent the m movement patterns of point P_i over the time interval as shown in Figure 4c. The first state represents the point staying still throughout the process. The other states represent $m-1$ movement patterns. For example, state S_3 represents that the point moves down and then comes back. State S_m shows a relatively more complex pattern in which the point moves down in the beginning and moves up later. These m states are mutually exclusive and encode the motion and direction information of each facial feature point. Movement features in the interval are collected and a k-means clustering is performed to determine the state of each primitive event. In conclusion, each facial feature point generates one primitive event. This event has m possible states and its duration is the time interval when the corresponding point is away from its neutral position.

The defined primitive facial events cover all of the local motions of the key facial components and provide us the basis to further study the spatio-temporal relations among them. They are explicitly obtained based on the tracking of the facial feature points and hence are easy to get without human labeling, training or prediction which could be time-consuming. Meanwhile, they also provide the time-interval information of the events and therefore allow us to study the relations of not only sequential but also overlapping events.

Given the time intervals of a pair of primitive facial events, we can then measure their temporal relation by calculating their temporal interval distance according to Equation 1 and Table 1. For instance, in Figure 4b the temporal relation is that E_2 overlaps E_1, with E_1 as the temporal reference. Figure 4d depicts the time intervals of a total of 26 facial feature points estimated from an image sequence in our experiment in which the expression is fear. From it we can clearly see the various temporal relations among the primitive facial events. The temporal relations will be evaluated for all the possible pairs of primitive facial events, but only those that exhibit high variance across different expressions will be maintained for expression recognition. This step is called temporal relation node selection and will be discussed in details in section 4.3.1.

4.2. Facial Expression Recognition

To recognize N facial expressions, we will build N ITBN's, with each corresponding to one expression. For



(a) **(b)** **(c)** **(d)**

Figure 4: (a) Facial muscle movement as captured by the movement of facial points. (b) Duration for event E_1 and E_2 and their temporal relation. (c) Typical movement patterns of a primitive facial event. (d) Time intervals for the primitive facial events.

each ITBN, the entity node represents the primitive event and the temporal relation node has K possible values, each of which corresponds to one temporal relation. Given a query sample x, its expression will be determined according to Equation 4, where M_y stands for the ITBN model of expression y. Since different ITBN models may have different spatial structures, the model likelihood $P(x|M_y)$ will be divided by the model complexity for balance. We use the total number of the links as the model complexity. The ITBN model that produces the highest likelihood will be selected.

$$y^* = \arg\max_{M_y} \frac{\log P(x|M_y)}{\text{Complexity}(M_y)} \tag{4}$$

4.3. Learning ITBN for Facial Expression

Learning the ITBN model for facial expression consists of three parts: temporal nodes selection, structure learning, and parameter estimation.

4.3.1 Temporal Relation Nodes Selection

While ITBN can capture the complex relations among the temporal entities, it is not necessary to consider the relation among all the possible pairs of events for facial expression recognition. A selection routine is hence performed to remove the pairs that may not contribute or may even do harm to expression recognition. With the goal of discriminating expressions, the relation between two temporal entities is expected to be strong and can maximally differentiate between different expressions. To meet this requirement, we define a KL divergence-based score to evaluate the relation node between each pair of events, and only retain those that have a relatively high score. The score of relation R_{AB} between event A and event B is defined in Equation 5, where P_i (P_j) is the conditional probability of R_{AB} for the i^{th} (j^{th}) expression with i (j) ranging over all the possible expressions. D_{KL} stands for the KL divergence. All the entity pairs are ranked according to their score. The top M pairs are selected and their temporal relations will be instantiated in the ITBN model.

$$S_{AB} = \sum_{i>j} (D_{KL}(P_i||P_j) + D_{KL}(P_j||P_i)) \tag{5}$$

4.3.2 Structure Learning

The next step is to learn the spatial and temporal links (i.e. the solid and dotted links in Figure 3) among the entity nodes and selected relation nodes. The temporal relation nodes can be directly linked to their corresponding events. Here we mainly focus on learning the spatial structure. Learning the ITBN structure means finding a network G that best matches the training dataset D. We use Bayesian information criterion (BIC) to evaluate each ITBN:

$$\max_G S(G:D) = \max_\Theta (\log P(D|G,\Theta) - |\Theta| \frac{\log N}{2}) \tag{6}$$

where S denotes the BIC score, Θ the vector of the estimated parameters, $\log P(D|G,\Theta)$ the log-likelihood function, and $|\Theta|$ the number of free parameters. The structure learning method proposed in [3] is employed to find the structure that has the highest BIC score.

4.3.3 Parameter Estimation

Parameters for ITBN involve the conditional probability distribution (CPD) for each node given its parents. Specifically, the conditional probability of each temporal relation node may have a large number of parameters since we have a large number of temporal relations and often don't have enough training data. To reduce the number of parameters to estimate, we employ a tree-structured CPD for each temporal node. An example is shown in Figure 5, which illustrates how we use a tree-structure CPD to parameterize the conditional probability of relation node I_{AB} given the event pair A and B. When A or B equals zero, meaning that they do not move, no information can be obtained about their temporal relation. Therefore the conditional probability is set to be uniform. When both of them move, the temporal relation probability holds regardless of their moving patterns. This parameterization method is specifically designed for insufficient training data, and does not limit us to use more complex CPD's if we have enough training samples.

Given a training dataset D which contains the properly estimated state of each primitive event and their temporal relations, the goal of parameter estimation is to find the maximum likelihood estimate (MLE) of the parameters Θ, which is shown in Equation 7. Θ denotes the parameter set and D represents the data.

$$\Theta^* = \arg\max_\Theta \log P(D|\Theta) \tag{7}$$

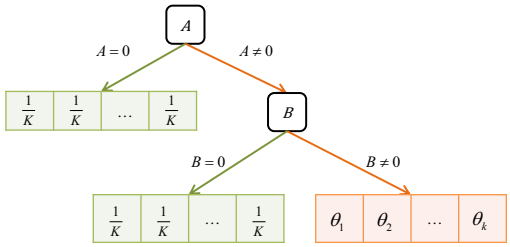


Figure 5: Tree Structured CPD for the Relation Node

5. Experiments

To evaluate ITBN, we study its performance on two widely used benchmark datasets, namely the extended Cohan-Kanade dataset [7, 6] and MMI dataset [12]. The goal is to evaluate if ITBN can improve performance by incorporating complex spatio-temporal relations, and to compare ITBN with the existing works.

5.1. Data

The extended Cohn-Kanade dataset (CK+) contains facial expression videos from 210 adults in which 69% are female, 81% are Euro-American, 13% are Afro-American and 6% are from other groups. Participants are 18 to 50 years of age. A total of 7 expressions are labeled in the dataset, including anger, contempt, disgust, fear, happy, sadness and surprise.

The MMI dataset includes more than 30 subjects in which 44% are female. The subjects age from 19 to 62 and are either European, Asian or South American. In this dataset, 213 sequences have been labeled with facial expressions, out of which 205 are with frontal face. Unlike other works that manually selected a subset of 96 image sequences for expression recognition, we use all 205 image sequences of 6 expressions from the MMI dataset in our experiment and perform recognition based on the image sequence without knowing the ground truth of the apex frames. Table 2 illustrates the number of samples for each expression in the two datasets.

Table 2: Number of Samples

Expression	CK+	MMI
Anger (An)	45	32
Contempt (Co)	18	
Disgust (Di)	59	31
Fear (Fe)	25	28
Happy (Ha)	69	42
Sadness (Sa)	28	32
Surprise (Su)	83	40

The two datasets present different challenges for facial expression recognition. All of the image sequences in CK+ start from the neutral face and end at the peak frame. Therefore, they only cover the first half of the expressions, which means for each event, we have its starting time and but not the end time. This effectively limits the temporal relationships to three relations which are A starts before B, A starts after B, and A starts at the same time as B. Image sequences in MMI cover the whole expression process from the onset to the offset. However, some subjects wear glasses, accessories or have mustaches, and there are greater intra-subject variations and head motions when performing expressions in MMI. These make it very difficult to analyze expressions.

The facial feature points that are used in our experiments are shown in Figure 6. For CK+, the facial feature points are provided by the database. For the MMI dataset, the facial feature points are obtained using an ASM model based method. The tracking results are normalized such that the eye centers fall on the given positions for all the frames based on affine transformation. To measure the duration of each event and deal with tracking noise, the point is said to move only when its relative distance from the neutral position exceeds 2 pixels.

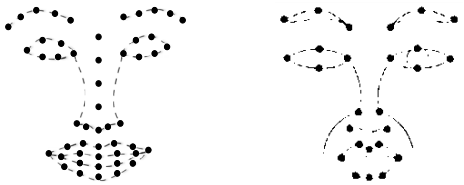


Figure 6: Facial Feature Points. Left: CK+; right: MMI.

Determining the moving pattern for each event requires collecting features during the motion interval. Since CK+ only covers half the process of the expression, we collect the moving directions during the motion interval and quantize them into four moving patterns. For MMI, moving features are collected as follows. We take the discrete Fourier transform of the moving trace of a point along both the horizontal and vertical direction and use the first 5 FFT coefficients as the feature. The direction of this point relative to its neutral position is collected for each frame and quantized into 4 orientations. A histogram of directions can be computed given the directions of all the frames during the event. All of these features are used to determine the state of the event by performing k-means clustering and a total of 9 patterns are used in MMI, including the stationary pattern. Experiments are performed based on 15-fold cross subject validation in CK+ and 20-fold cross subject validation in MMI.

5.2. Performance Vs Number of Relation Nodes

The first experiment is to evaluate if incorporating temporal relations could enhance the performance of facial expression recognition. Since not the relation of all the event pairs will be helpful for expression recognition, we performed a selection subroutine and picked those that have relatively high scores. In this section we evaluate the performance with respect to the number of temporal relation nodes we selected. Figure 7 illustrate the performance of ITBN in CK+ and MMI when we gradually increase the number of relation nodes. The x axis represents the num-

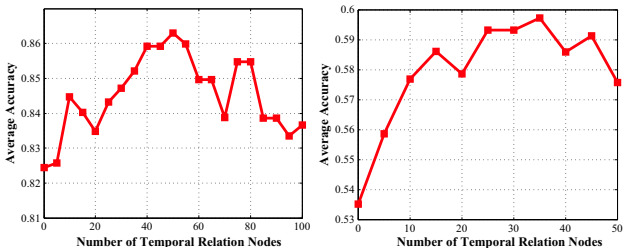


Figure 7: Performance of ITBN with respect to the num of event pairs in CK+ dataset (left) and MMI dataset (right).

ber of temporal relation nodes we selected. The average recognition rate is calculated by averaging the classification accuracy for each expression and corresponds to the y axis. The starting point of each curve is the performance of the model when no temporal relation is incorporated. From the results we can see that by incorporating the temporal relation information, the recognition rates of both models are significantly improved. The performance reaches its peak when approximately the 50 events pairs for CK+ and 35 pairs for MMI are selected. In CK+ the recognition improvement is about 4% and in MMI the improvement is about 6%. This demonstrates the benefits of temporal information for expression recognition and the ability of ITBN to capture such knowledge. As more and more relation nodes are added to ITBN, the performance will eventually decline. This is partially because the contributions of the low-score relation nodes to classification could be less than the noise they bring. Table 3 and Table 4 show the confusion matrices of ITBN in two datasets when 50 event pairs for CK+ and 35 pairs for MMI were selected. The corresponding average recognition rates of these two matrices are 86.3% and 59.7%.

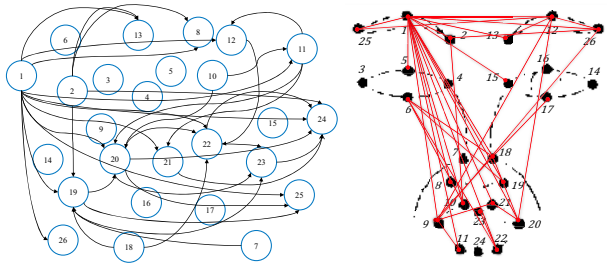
Table 3: Confusion Matrix of ITBN in CK+

	An	Di	Fe	Ha	Sa	Su	Co
An	91.1	0.0	0.0	4.4	0.0	2.2	2.0
Di	1.2	94.0	1.2	0.0	1.2	2.4	0.0
Fe	5.6	0.0	83.3	0.0	0.0	0.0	11.1
Ha	3.4	0.0	0.0	89.8	1.7	3.4	1.7
Sa	0.0	20.0	0.0	0.0	76.0	4.0	0.0
Su	0.0	5.8	1.5	0.0	0.0	91.3	1.5
Co	7.1	0.0	3.6	0.0	10.7	0.0	78.6

Table 4: Confusion Matrix of ITBN in MMI

	An	Di	Fe	Ha	Sa	Su
An	46.9	18.8	0.0	3.1	31.2	0.0
Di	16.1	54.8	9.7	6.5	6.5	6.5
Fe	7.1	10.7	57.1	10.7	3.6	10.7
Ha	0.0	7.1	19.1	71.4	2.4	0.0
Sa	9.4	3.1	18.8	3.1	65.6	0.0
Su	0.0	2.5	32.5	2.5	0.0	62.5

Figure 8a illustrates the learned ITBN model in MMI dataset, where each node represents an event corresponding to a facial feature point, and each link denotes a pair-wise temporal dependency. To gain some insight of the temporal interactions of the facial muscles, Figure 8b graphically depicts all of the 35 selected temporal relation nodes in the MMI dataset. If the relation node R_{AB} is selected, then a line is connected between event pair A and B. In particular, the frequencies of all the thirteen relations between feature point 1 and 11 are shown in Figure 9. We can see that selected interactions provide discriminative information to recognize expressions and they involve all components of the face.



(a) **(b)**

Figure 8: (a) The learned ITBN model in MMI dataset. (b) Graphical depiction of the selected event pairs in MMI.

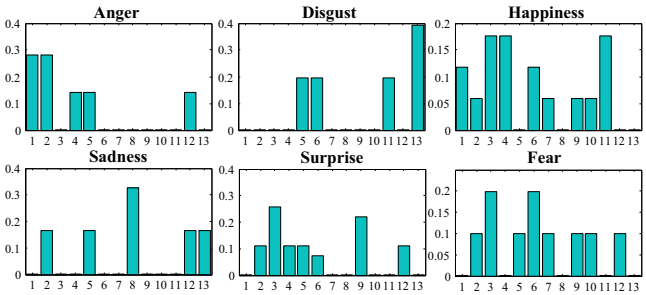


Figure 9: Frequencies of thirteen relations among a pair of events with respect to different expressions in MMI. X-axis represents the index of relationships.

5.3. Comparison with Related Works

From the previous experiment we can see that ITBN can successfully capture and exploit the spatio-temporal information to enhance expression recognition. In this section we compare the performance of ITBN with related works. We also evaluate ITBN against other time-sliced dynamic models. Specifically a hidden Markov model (HMM) which is based on the locations of facial feature points is implemented, and we expect similar results for the DBN model.

Our experiment faces more challenges than those represented in many other works. First, we perform recognition on a given sequence without knowing the ground truth of the peak frame. Secondly, our model only uses the tracking results without any texture features such as LBP or Gabor. This makes it more difficult for us to recognize expressions. Furthermore, in the MMI dataset, we use all of the 205 image sequences instead of manually selecting 96 sequences for recognition.

Table 5 compares the result of ITBN with that in [7] where they use the similarity normalized shape fea-

tures (SPTS) and canonical normalized appearance features (CAPP) that are computed based on the tracking results of 68 facial feature points. We can see that ITBN outperforms [7] by about 3%.

Few works can be found that use tracking results for expression recognition in MMI dataset. Among all the works we can find, [15] is the most similar to ours in that they also use all of the 205 sequences. Their method is based on the LBP features and they propose to learn the common and specific patches for classification. Table 6 shows both our and their results, in which CPL stands for their method that only uses common patches, CSPL is their method that uses common and specific patches and ADL is the patches that are selected by AdaBoost. We can see that our results are much better than CPL and ADL. Although CSPL outperforms our result, their experiment is based on appearance features and requires the peak frames while we only use the features from the tracking results and do not have the ground truth of peak frame.

On both datasets, the results of HMM are also illustrated in the above two tables. During the experiment, we chose 4 and 10 latent states for HMM in CK+ and MMI respectively such that the recognition rate of HMM is maximized. ITBN outperforms HMM in both cases.

Table 5: Comparison in CK+

Method	ITBN	HMM	Lucey *et al.* [7]
AR %	86.3	83.5	83.3

Table 6: Comparison in MMI

Method	CPL	CSPL	ADL	ITBN	HMM
AR %	49.4	73.5	47.8	59.7	51.5

Overall we can see that ITBN can successfully capture the complex temporal relations and translate them into the significant improvement of facial expression recognition. ITBN outperforms the time-sliced dynamic models and other works that also use tracking-based features and can achieve comparable and even better results than those appearance-based approaches.

6. Conclusions

In this paper we model a facial expression as a complex activity that consists of temporally overlapping or sequential primitive facial events. More importantly, we have proposed a probabilistic approach that integrates Allen's temporal Interval Algebra with Bayesian Network to fully exploit the spatial and temporal interactions among the primitive facial events for expression recognition. Experiments on the benchmark datasets demonstrate the power of the proposed method in exploiting complex relations compared to the existing dynamic models as well as its advantages over the existing methods, even though it is purely based on facial feature movements without using any appearance information. Moreover, ITBN is not limited to model relations among the primitive facial events and could be widely applicable for analyzing other complex activities.

Acknowledgement

This work is jointly supported by an NSF grant (#1145152), a US Army Research Office grant (W911NF-12-1-0473), and an NSFC grant (61228304).

References

[1] J. F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, 1983. 3

[2] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang. Facial expression recognition from video sequences: Temporal and static modelling. In *Computer Vision and Image Understanding*, pages 160–187, 2003. 1, 2

[3] C. P. de Campos, Z. Zeng, and Q. Ji. Structure learning of Bayesian networks using constraints. In *ICML*, 2009. 5

[4] S. Jain, C. Hu, and J. K. Aggarwal. Facial expression recognition with temporal modeling of shapes. In *ICCV Workshops*, pages 1642–1649. IEEE, 2011. 2

[5] R. E. Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *CVPR Workshop*, 2004. 1, 2

[6] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46 –53, 2000. 6

[7] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshop*, pages 94–101, june 2010. 6, 7, 8

[8] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, 1998. 2

[9] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971 –987, jul 2002. 2

[10] M. Pantic and L. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 34(3):1449 –1461, june 2004. 1

[11] L. Shang and K.-P. Chan. Nonparametric discriminant hmm and application to facial expression recognition. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 2090–2096, June 2009. 1, 2

[12] M. F. Valstar and M. Pantic. Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database. In *Proceedings of Int'l Conf. Language Resources and Evaluation, Workshop on EMOTION*, pages 65–70, Malta, May 2010. 6

[13] M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, pages 28–43, 2012. 1, 2

[14] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, Jan 2009. 2

[15] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. Metaxas. Learning active facial patches for expression analysis. In *CVPR*, pages 2562–2569, june 2012. 8